

Crowdsourcing Documentary Transcription: an Open Source Tool

I. Purpose and long-term goals

A. Overview

The NHPRC funded *Papers of the War Department, 1784-1800* project <wardepartmentpapers.org> is a cutting edge digital archive that makes available high resolution images and basic metadata for nearly 45,000 documents. The staff of editors at the Center for History and New Media (CHNM) at George Mason University has been engaged in processing these documents since 2006, and we are on track to eventually provide extended keyword metadata. However, the sophisticated and powerful search engine is only as powerful as the data it can index. Without transcriptions of the documents, users will never be able to access the full potential of this archive. Unfortunately, the project will never have the resources to pay for the staff to transcribe the documents. This is common problem for digital archives in an era of scarce resources. Luckily, the transcription process is ripe for innovation. **Thus, CHNM requests \$69,145 over eighteen months from the National Historical Publications and Records Commission (Publishing Historical Records) to build and distribute an open source tool to enable researchers to contribute document transcriptions and research notes to the *Papers of the War Department* (PWD) website.** This tool will be fully generalizable so that other digital documentary projects can join PWD in harnessing the power of the community of users to improve the discoverability and usefulness of their archives.

B. Background and Assessment of Need

On the night of November 8, 1800, fire devastated the War Office, consuming the papers, records, and books stored there. Two weeks later, Secretary of War Samuel Dexter lamented in a letter that “All the papers in my office [have] been destroyed.” From the perspective of historians, the loss was monumental since in many respects the documents lost in the fire constituted the first “national archive” of the United States. For two hundred years, military historians virtually ignored the decades prior to 1800 because they had so few primary documents available to examine. The *Papers of the War Department, 1784-1800* (PWD) is an ongoing digital editorial project that encourages historical scholarship of this lost period in the Early Republic by restoring the archive online and making the entire collection accessible to a wide audience of scholars, students, teachers, and the general public.

CHNM took over the work on PWD in 2006, building on the efforts of Ted Crackel (formerly of East Stroudsburg University and now Editor-in-Chief of the George Washington Papers at University of Virginia) to recreate the archive with scanned images from repositories around the country. With funding support from NHPRC and the National Endowment for the Humanities we built a robust web architecture and searchable database to serve the documents. This website <wardepartmentpapers.org> officially debuted in June 2008 with nearly 45,000 document images and metadata. The site is on target to receive 49,000 unique visitors in 2009 with minimal outreach and publicity.

Even the best search for a large archive is limited by the metadata that it can access. Thus, a scholar searching the PWD archive for documents that include the keywords “Indian,” “treaty,” “negotiation,” “boundary,” and other terms will easily find a May 2, 1788 letter from Arthur St. Clair to Charles Thomson discussing negotiations with Indian tribes. But since the PWD archive contains only basic keyword metadata, rather than full transcripts of the documents, the engine cannot rank the search results according to relevance and significance. A visitor performing that keyword search

will likely receive more than 150 items along with the St. Clair-Thomson letter. Many of those documents are brief, mentioning the terms only once, and are therefore likely to be of less interest to researchers. Without a full transcription, the search engine cannot distinguish between these short items and the far more relevant letter, which mentions each term dozens of times over its five pages.

Realistically, we cannot expect sufficient funding to support an army of staff and associate editors to centrally transcribe each document of this vast archive, with its often difficult to read eighteenth century handwriting. As an assembled archive, rather than the papers of a single individual, the array of authors and handwriting idiosyncracies mean that automated handwriting recognition is not an option. Even if it were, that technology is not nearly advanced enough to produce an error level small enough to function accurately with search.¹ Thus, we are left with the problem of how to make this tremendously important historical collection more accessible to researchers when we cannot hope to undertake the task of in-house hand-keyed transcription.

Of course, this is not a problem that is limited to PWD. Many efforts to publish the historical documents of the nation's founders have been decades in the making and still have significant volumes yet unpublished. For example, the Papers of George Washington project began in 1969 and has published 52 of 90 expected volumes in the past 40 years. On a smaller scale, the Papers of James Madison project, which represents an archive of roughly 27,000 documents, began in 1956 and has not yet published one-third of the expected volumes (15 of the 45).² This important work has taken hundreds of thousands of hours of dedicated staff time and millions of dollars in funding. Proceeding chronologically, documentary editing projects may face financial hardships before they transcribe and publish the heart of the historically significant elements in a collection.

¹ Réjean Plamondon, Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1 (January 2000): 63-84.

² See, "Project and Volumes," *The Papers of George Washington* < <http://gwpapers.virginia.edu/project/index.html> > and "About," *The Papers of James Madison* < <http://www.virginia.edu/pjm/description1.htm> >.

The creation of digital documentary archives can provide the advantage of making archival material available to the public with relative quickness, before complete metadata has been produced. Yet this non-traditional publication schedule also carries with it certain drawbacks. Even with this acceleration, transcription of handwritten documents remains a final hurdle. For both traditional print editions and digital editions, the transcription process requires a commitment of tremendous amount of project resources, both temporal and fiscal. Though transcription is but one element of documentary editing, it is an essential one, and one that is ripe for innovation. Digital archives and documentary projects need a viable solution that lowers both the cost and the investment of staff time involved with transcribing of large numbers of historical documents. This grant seeks to address these long-term resource challenges facing many documentary editing projects.

C. An Opportunity to Capitalize on the Wisdom of Users

While the staff at the PWD will never have the capacity to transcribe a significant number of the documents in the archive, our yearly site traffic suggests that we have a vast resource of scholars and researchers who can help with this task. Everyday researchers examine documents in the archive, regularly making rough transcriptions to use in their own work. We routinely get emails from researchers suggesting improvements to the archive's metadata based on their work with the materials. **Thus, we propose to build an open source transcription tool to allow users easily to submit those transcriptions and that knowledge back to the archive.** The resultant tool will allow PWD, and eventually other digital archival projects, including examples from Minnesota Historical Society and the Independence Seaport Museum that will serve as test sites, to draw upon the wisdom of the thousands of interested researchers, scholars, and students who work with these materials (*See Appendix A: Letters of Commitment*). Gradually, users' combined work will enhance the

discoverability and usefulness of the archive without significantly adding to the costs of the project. We envision this tool as a response to Max Evan's 2007 call for commons-based peer production as way to create "Archives of the People, by the People, for the People."³

In proposing to open the transcription process up to crowdsourcing, CHNM is drawing upon the example and success of a host of other successful ventures. Though the goals and needs of digital archive and documentary editing projects are distinct from the goals of these "crowdsourcing" ventures, they do offer promising glimpses of the efficiencies and outcomes for our purposes.⁴ Each taps into the interests and passions of a segment of the public who contribute to the accumulated value and content of a project—whether software development or knowledge aggregation. The open source software movement—with successes like the Linux operating system and Mozilla's increasingly popular Firefox web browser—contains lessons for those of us interested in harnessing the expertise of a particular community to build a successful project. The developers who participate in open source software projects do so for their own reasons, but they contribute to the common good by applying their expertise to the demands and problems raised by software innovation. There are, however, several content-focused examples of crowdsourcing that are particularly revealing.

Wikipedia <en.wikipedia.org>, the free online encyclopedia that is written and maintained by users around the world, is by far the most well-known instance of successful crowdsourcing. Launched in 2001, using a simple authoring and versioning software to manage articles, Wikipedia thrives on the contribution of thousands of anonymous users. The site currently receives roughly 65 million visitors each month and has more than 75,000 active contributors who create and edit over

³ Max J. Evans. "Archives of the People, by the People, for the People," *American Archivist* 70:2 (2007): 387-400.

⁴ The term "crowdsourcing" was coined by Jeff Howe in "The Rise of Crowds," *Wired* 14:06 (June 2006), available at <<http://www.wired.com/wired/archive/14.06/crowds.html>>.

13 million articles—more than three million of which are in English.⁵ From its founding, teachers, parents and scholars have worried about the accuracy and content of Wikipedia, especially since articles from the free encyclopedia have long been the first result to appear in most search engines. Yet, the source of the occasional errors and incoherence of Wikipedia is also its power—the tremendously fast-moving contributions and collaboration of users means that errors are unlikely to remain for long before they are corrected. Similarly, the growth of articles and topics covered in the free encyclopedia reflects the energy and interests of the open source community of users. An assessment of the creation and revision history of articles on historical topics could provide a tremendous wealth of information about the concerns of the public with the past.⁶

In a similar, albeit much less extensive way, Flickr Commons <flickr.com/commons/> provides another example of tapping the collective knowledge of interested members of the public. Beginning with a seed contribution by the Library of Congress in January 2008, cultural institutions from around the world have contributed images and associated metadata to the Flickr Commons collection including the New York Public Library, the George Eastman museum, the Powerhouse Museum in Australia, and the Smithsonian Institution. The collections are freely available to the public, who can tag and comment on each image. A year ago, the Library of Congress assessed their participation in the pilot project, reporting that over 2,500 Flickr users had left more than 7,000 comments on almost 3,000 images. Library staff culled important corrections and additions to captions, titles and the identification of individuals that were then incorporated back into the metadata of more than 500 items by August 2008. Based on these positive interactions with the public, the staff advocated that the library continue to draw upon public knowledge through the

⁵ “About,” Wikipedia <http://en.wikipedia.org/wiki/About_Wikipedia>.

⁶ Roy Rosenzweig, “Can History be Open Source? Wikipedia and the Future of the Past,” *The Journal of American History* 93:1 (June, 2006): 117-46, available at <<http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=42>>.

Flickr Commons project and other Web 2.0 interactive projects, declaring: “The benefits appear to far outweigh the costs and risks.”⁷ The Library of Congress's success helped to encourage the Smithsonian Institution to join Flickr Commons, as a result, in the period between June and December 2008, their photographic contributions received over 625,000 views with many user comments and tags. Thus, trusted cultural heritage organizations are beginning to recognize the powerful ways that interested members of the public, scholars, and educators can contribute to the knowledge base related to collections.

To bring this type of crowdsourcing to the world of digital archives, CHNM will create a transcription tool that makes it easy for users to contribute transcription text for the documents they read and use in their research. The faithful and literal transcription of a group of documents with many authors is a significant problem that requires constant reflection on and examination of the images of such documents. The design and display innovations provided on the internet today make this process of examination and re-examination accessible to thousands of scholars. In the case of the PWD, a researcher might be working with the May 2, 1788 letter from Arthur St. Clair to Charles Thomson that discusses negotiations with Indian tribes. Locating the record in the digital archive, a researcher currently has access to the metadata for the letter and a link to a high resolution image of each page of the document. With the addition of the transcription tool, a user will also have the option to contribute a transcription of the document, to engage in discussion of the item, and to review the editing history of the transcription, if there is one.

Should one of these users click on the link to contribute a transcription, she will be brought to a new webpage consisting of two panes. On the left side of the page, a document viewer will provide access to the high resolution images of the letter, with the capacity to zoom in on the images

⁷ Michelle Springer, et. al., *For the Common Good: The Library of Congress Flickr Pilot Project, Final Report* (October 30, 2008): 36, available at <http://www.loc.gov/rr/print/flickr_report_final.pdf>; Martin Kalfatovic, et al. “Smithsonian Team Flickr: a library, archives, and museums collaboration in web 2.0 space,” *Archival Science* (October 2009), available at <<http://dx.doi.org/10.1007/s10502-009-9089-y>>.

and page back and forth between them if there are several. The right pane of the screen will include a login interface for first time users. This login and authentication system will protect the archive from vandalism. Once the user has logged in, she will be presented with a simple wiki editing window into which she can enter her transcription of the document, drawing on the formatting conventions available through MediaWiki, which will be familiar to anyone who has edited a Wikipedia page. If another researcher has already worked on this particular document, that transcription and mark-up will reside within the editing window. Additionally, the saved version of the transcript will be printed below the editing window. Once the user is done keying in her transcription, she will click to save the text. Then, if she returns to the original document record within the archive, as the most recent contributor, her version of the transcription will appear with the letter's metadata. If the user has significant knowledge about the St. Clair-Thomson letter, upon returning to the main item record, she could click on the discussion link to leave some notes about the document. Similarly, if she were unsure about some element of her transcription, she could note that for future users. (*See Appendix B: Mock-ups.*)

Unlike traditional crowdsourcing ventures, opening transcription of digital archives will necessarily draw upon the contributions of a specialized audience of users. The user base for the PWD is composed primarily of scholars interested in the Early Republic, military, and Native American history—advanced undergraduate students, graduate students, junior scholars, and senior scholars, all of whom have their own reasons for creating transcriptions of particular documents. Certainly there are also members of the interested general public who use the archive in their work to research family history and other related topics. The addition of the transcription tool to the PWD digital archive will allow us to reach out to an entirely new audience that will include documentary editors in training. The open interface will allow these students to operate on a live archive, practicing the techniques and elements of faithful archival transcription that they are

learning in their classes and seminars. As other digital archives begin to implement the tool, the array of document types and conditions on which proto-editors can practice will grow. Together these users compose a specialized audience who can contribute both their transcriptions and insights to the collective knowledge base of the archive.

Public transcription of digital archival material will not immediately produce perfect results. However, just as the errors in Wikipedia are quickly corrected by other users, the open access provided by the transcription tool will allow the editors and other users to correct mistakes as they become apparent. PWD's Assistant Editors will provide oversight, reviewing document discussions, mining research notes to add to existing metadata and cross referencing.

This scenario suggests that there will be significant benefits for interested users, whether they are scholarly researchers, students and teachers, or members of the interested public. First, general users will immediately benefit from the improved search capacity that full-text transcriptions will yield. Second, students of documentary editing and their instructors will have access to a space where they can participate in the process of transcription while dealing with actual archival sources. Students will have the experience of approaching the documents as editors and contributing to the work of the project. Third, the modified wiki interface will allow users to engage in an online conversation about particular documents. Researchers will be able to leave notes for the editors, both about the document transcription, but also about the item's metadata and its relationship to other items in the archive. Thus, the public transcription tool will create a venue for scholarly communication, collaboration, and community building based on the documents.

Similarly, crowdsourcing transcription for digital archives will provide significant benefits to documentary editors. First, and foremost, public contributions will provide transcriptions where there once were none, and where there likely would be none in the future. Moreover, archival staff will be able to draw on the publicly contributed transcriptions to form a base for their editorial

work. Editors may start with a rough transcription provided by interested users, and then apply the techniques and expertise of their training to produce a corrected transcription quickly. Moreover, allowing the public to contribute document transcriptions to digital archival collections will have real-time benefits for the accessibility of the documents. Each transcription contributed to the archive will immediately be available to the collection management system's search engine. The result will be ever-improving discoverability. The full text of the documents will allow the search engines to surface the most relevant documents by better weighting their results. Also, crowdsourcing transcriptions will allow editors to better understand the ways in which users interact with their archive. Necessarily the documents that are of the most interest to users will be transcribed most quickly and fully. Thus, the public contributions can serve as a barometer of the most interesting materials within a particular collection.

In sum, crowdsourcing transcription has the potential to change what is possible in the world of digital documentary editions by facilitating faster, more complete, useful work drawing on fewer resources. For example, in testing the tool, the Minnesota Historical Society is considering making public documents from several different collections: the 102 page Alexander G. Huggins diary, (1835-1846); selections from Ernest Oberholtzer's canoe trip journals documenting his trips and Ojibwa people and cultures; or selected items from the James J. Hill collection and letterpress books that deal with the reorganization of the Northern Pacific railroad. Independence Seaport Museum, on the other hand, is considering testing the tool with archival collections pertaining to American trade with China during the late 18th and early 19th centuries. Once transcribed, sources from any of these collections will open up new avenues of inquiry for scholars who will be able to easily locate the documents that relate to their research questions.

II. Plan of work

Our work on the crowdsourcing transcription tool will span 18 months and will be composed of six phases. Each phase will involve a significant commitment of staff time, as indicated in the budget. (*See also, Budget Narrative and Appendix C: Work Plan Chart.*):

- **July 2010-September 2010: Planning:** During the first three months of the grant, we will do intensive planning work using the PWD website as a case study. The project team will create a set of wire frames and user interface scenarios that will help predict the best workflow for moving users from the existing digital archive to the transcription interface. Also, we will establish best methods for using MediaWiki as the database and formatting engine behind the transcription tool. This will require a thorough knowledge of the MediaWiki application programming interface (API). The result of this phase of work will be a set of workflow diagrams and schema that will guide the production of the transcription tool.
- **October 2010-June 2011: Production of the transcription tool:** After the initial planning stage, the web development team will embark upon building the transcription tool and adapting it for the PWD website. This will represent the bulk of the technical work for the project. Creating the tool will require two steps (*See Appendix B: Mockups*):
 1. We will build a generalizable tool for document transcription, comprising an Web interface powered by a lightweight PHP library designed to interface the MediaWiki API and any standards compliant content management system (CMS). Though we considered building a set of plug-ins to connect the transcription tool to particular digital archival systems, we decided in favor of creating a PHP library because it will make our tool completely generalizable and adaptable.

- The **Web interface** will have two major components: a document viewer and an editing window for the transcript input. The **JavaScript document viewer** will make it possible for users to page through all of the images that accompany a multi-page document, and to zoom in on high resolution image so that they can better decipher difficult handwriting and marginalia.. The transcription form will include a login interface, and editing window and a link to helper screen that summarizes the text formatting conventions
 - The **PHP library** will have two major components. The MediaWiki API client will communicate with the MediaWiki API server. The CMS adapter interface will define the requisite data needed from the CMS for the transcription tool to function properly.
2. We will integrate the transcription tool into the PWD archive by writing an **adapter** for the tool and an **API server for PWD**, so that we will be able to access the high resolution images for each document via an XML or JSON feed. We will write a script that uploads the most current version of the transcription into the PWD database, allowing us to publish and search the most up-to-date version of each document transcription. Thus, test developers will be able to write a similar adapter tailored to their specific CMS (e.g. Omeka, Drupal, Wordpress, Joomla, custom systems).
- **July 2011-December 2011: Implementation:** Once we have built the elements of the transcription tool, we will distribute a beta version to our testing partners, Minnesota Historical Society and Independence Seaport Museum (*See Appendix A: Letters of Commitment*). Also, the beta code will be made available through Google Code under the open source General Public License (GPL), so that any developer can access and modify the

tool. We will provide our testers with two levels of support as they use the transcription tool with their collections:

1. We will hold an introductory video conference to connect our web development team with the programmers and developers from the test sites. This conference will include a discussion of customizing and adapting the the transcription tool to work with various digital archive content management systems.
 2. Our web development team will hold “office hours” in IRC Chat on Fridays during the implementation and testing phase of the project, so that test developers can have ongoing access to advice and help as they work to install and customize the tool for their digital archives.
- **July 2011-December 2011: Evaluation:** We will conduct several layers of qualitative evaluation of the transcription tool. For our official testing partners, we will create a survey that covers several areas of inquiry, including the quality of the experience with the user interface for editors, the function and stability of the tool's code, and other features that they might find useful. In addition to surveying our testers, we will create a voluntary survey for researchers and archive users, to gather information on their experience in the process of contributing transcriptions to the archive. Together, this qualitative information will combine with quantitative information gathered from server logs on page visits and the cumulative total of transcription edits, user accounts created, and documents with discussion to form a picture of the potential for crowdsourcing digital archival transcription. After gathering this feedback from our testers and users, we will make any necessary changes in the functionality before we officially release and publicize for the tool.
 - **July 2011-December 2011: Authoring the *Guide to Crowdsourcing Documentary Transcription*:** Upon completing the construction of the elements of the transcription tool

and implementing it with the PWD, the project team will embark upon writing a guide for other editors who wish to implement the transcription tool with a digital archive. The guide will address two main concerns: the technical and the practical. On the technical side, we will provide explicit documentation on implementing and customizing the tool for use with a variety of collection management systems, including tips on creating an adapter plug-in for various platforms, skinning the tool to match a website design, and modifying the user account login and permissions structures. On the practical side, we will provide advice on encouraging contributions from the research community, and the demands managing the oversight of transcriptions and document discussion notes by digital archive editors. We will make the *Guide* freely available as text and PDF on the site that we create to distribute the transcription tool.

- **July 2011-December 2011: Dissemination:** After implementing any necessary revisions to the transcription tool, we will make an official release of the tool and the *Guide*. (*See Dissemination below for a full description of our distribution and publicity plan.*)

III. Products and Dissemination

A. Products:

- An open-source transcription tool that brings together 1) a modified wiki package with versioning, user login and authentication, and the ability for discussion; 2) a JavaScript-based document viewer that allows users to zoom in on high resolution images of documents.
- A model for creating CMS adapters that will connect the transcription tool to potentially any content management system (Omeka, Drupal, Wordpress, etc.), allowing both for the call of the document images and the return of the transcription text to the main archive.

- *A Guide to Crowdsourcing Documentary Transcription* that provides both technical and practical advice for editors of digital documentary archives who wish to deploy the tool.

B. Dissemination

We will widely disseminate our work both to the world of documentary editors and the world of digital humanities, through the *CHNM Labs* portal <chnm.gmu.edu/labs/>. The CHNM website receives significant traffic each year, accruing a total of over 11 million unique visitors in 2008. The products of this grant will fit nicely with the other cutting edge projects housed at *CHNM Labs*, such as *Mobile for Museums* <chnm.gmu.edu/labs/mobile-for-museums/> a project funded by the Samuel H. Kress Foundation to study the use of mobile devices in the museum world and to develop several innovative prototypes to improve that usage. From *CHNM Labs*, we will link to the source code, publish the PHP library, and the *Guide to Crowdsourcing Documentary Transcription*. Since the transcription tool is an open source project, we will publish the source code at Google Code <code.google.com> under a General Public License (GPL).⁸

Simultaneously, we will launch a broad-based publicity campaign. CHNM sponsors a number of widely read professional blogs, and these blogs have proven to be a very effective means for rapidly disseminating news to those interested in digital tools. We will use them to reach out to the digital humanities community. To encourage researchers to participate in the PWD project, we will publicize the additional functionality on the archive's website and post announcements to relevant listservs to attract our core audience of scholars (e.g. H-SHEAR, H-OIEAHC H-FedHist, H-AmInd, H-War). In reaching out to a more general audience, we will also make use of the History News Network, a CHNM affiliate that three times per week sends a newsletter to 13,000

⁸ Free Software Foundation, "GNU General Public License," <<http://www.gnu.org/licenses/gpl-3.0.html>>.

scholars, educators, public historians and members of the interested public who may be interested in participating in the work of transcription. Additionally, we will reach out to documentary editors who may wish to use the tool with their archives by publicizing it with important discussion networks (e.g. SEDIT-L), and by making presentations at the Association for Documentary Editing and the Society of American Archivists' annual meetings. Finally, we will tap into the institutional networks of our testing partners to reach small archives and museum that would be interested in using the tool.

IV. Qualifications of the staff

CHNM's extensive experience in digital projects equips it to manage the planning, development, and dissemination of this public transcription tool. The project builds directly on three core strengths of CHNM: 1) our development of widely used open-source digital tools, including *Omeka* (recent winner of a Mellon Award for Technology Collaboration) and *Zotero* (named among *PC Magazine's* "Best Free Software" two years running); 2) our creation of large-scale, preservation-grade digital archives, such as the *September 11 Digital Archive* <911digitalarchive.org> and the *Hurricane Digital Memory Bank* <hurricanearchive.org>, which have fostered an interactive model for documenting and preserving the past; 3) our use of the web for more than a dozen years as a vehicle for engaging broad and diverse audiences, often in partnership with libraries, archives, and museums, for example on our NEH-funded *Bracero History Archive* <braceroarchive.org> project (a partnership with the National Museum of American History, Brown University and the University of Texas, El Paso) or on *Martha Washington: a Life* <marthawashington.us> (a partnership with the Mount Vernon Estate and Garden). This long-standing track record of outstanding and innovative work in digital history projects makes us the ideal candidates to build and test an application for crowdsourcing document transcription. (See *Appendix D: Organizational Profile*.)

Moreover, we have assembled a uniquely capable staff to build this tool and test it with the PWD. Editor-in-Chief Christopher Hamner has overseen the project since its arrival at Mason in the spring of 2006. His ongoing oversight of the editorial staff will ensure that the public contributions to the website will be carefully monitored and assessed for their value to the project. Sharon Leon (Director of Public Projects, CHNM) has overseen work on numerous digital archival and cultural heritage projects drawing on the capacities of new technologies to transform the field of history. Together, Hamner and Leon will oversee the work of Jim Safley and Chris Raymond as they design and construct the transcription application and the user interface, and the work of the Assistant Editors as they monitor and evaluate the public contributions. (*See Participants' Resumes.*)

Key Personnel:

Christopher Hamner (Co-Primary Investigator and Editor-in-Chief) has been responsible for overall editorial content and intellectual direction of the project, supervising the Assistant Editors, and overseeing the day-to-day editorial work of the project since 2006. Hamner earned his PhD in American military history from the University of North Carolina, where he studied with Richard Kohn. He is a member of the Association for Documentary Editing and attended NHPRC's 2007 Institute for the Editing of Historical Documents. He is completing a book that examines the ways that American infantry soldiers responded to the traumas of combat in the wars of two centuries, beginning with the War for Independence.

Sharon Leon (Co-Primary Investigator and Project Director) will assist in supervising the project and will coordinate its technical aspects. Leon is Research Assistant Professor of History at George Mason University and Director of Public Projects at the Center for History and New Media. She has a bachelor's degree from Georgetown University and a doctoral degree in American Studies from the University of Minnesota. Her book *An Image of God: Catholics and American Eugenics* is forthcoming from the University of Chicago Press. At the Center, Leon oversees the projects that

involve work with museums, archives, and cultural heritage institutions. She also co-directs *Omeka* <omeka.org>, an open-source collections based web-based publishing platform for scholars, librarians, archivists, museum professionals, educators, and cultural enthusiasts.

James Safley (Technical Director), the architect of the Papers of the War Department's database and the designer of its powerful and flexible search engine. Safley received his undergraduate degree in history at GMU. He began his archiving career in 1999 at the National Archives and Records Administration, and continued that work at Phi Beta Kappa and at Mason's Special Collections and Archives. Coming to CHNM in 2002, Safley applied his traditional archiving experience to his work in digital archiving, web programming, and database administration. He will do the work of programming the tool, creating the PHP library, and advising the technical staff of our testing partners.

Chris Raymond (Web Designer) has diverse experience creating effective, award-winning visual communications as a graphic designer, web designer, science museum exhibit researcher/writer, science journalist, and publications director for an association of hands-on science museums. She has a Ph.D. in sociology from Cornell University with emphasis in mass media and sociology of science. Raymond formally studied graphic design at Virginia Commonwealth University, and specializes in typography, color, integrating imagery and text, and information hierarchy. She will create and style the user interface for the tool, including the image viewer.

Finally, the **Assistant Editors** will monitor the public contributions to the archive. Currently, they do the day-to-day work with the documents themselves. PWD currently employs five part-time Assistant Editors draw from American history students in Mason's History Ph.D. program who together handle the complex work of preparing document metadata. Reading and processing the documents accurately and efficiently requires knowledge of the Early Republic, training in history, and sensitivity to the nuances of eighteenth-century language. The current staff of

Assistant Editors includes collectively boast nine years of experience working on the project. The wages included in the proposal budget will allow for an Assistant Editor to provide oversight of the public interaction for PWD.

V. Objectives

- To build an open source transcription tool that will allow users to contribute work to digital documentary editions
- To build a CMS adapter interface that allows that tool to function with standards compliant system
- To implement the tool with the *Papers of the War Department* project
- To engage members of the user community in contribution transcriptions to PWD
- To test the tool with two project partners
- To author a *Guide on Crowdsourcing Documentary Transcription* that will provide both technical and practical advise to editors
- To disseminate the open source tool widely to digital archivists